Complex Knowledge Base Tools for Natural Language Communication with the Semantic Web

Kamil Ekštein[1], Dalibor Fiala[1], Miloslav Konopík[1], Roman Mouček[1], Tomáš Pavelka[1], Josef Steinberger[1], Roman Tesař[1], and Michal Toman[1]

University of West Bohemia, FAS, DCSE
Univerzitní 8, 306 14 Pilsen, Czech Republic
konopik@kiv.zcu.cz

Abstract. In this article we introduce the COT-SEWing (Complex Knowledge Base Tools for Natural Language Communication with the Semantic Web) project. The purpose of the project is to develop a complex base of tools which will remove some of the typical barriers present in communication between human user and computer within the scope of internet access. Our contribution includes methods, algorithms, tools and techniques improving the quality of user's experience during web usage.

1 Introduction

In this article we introduce the COT-SEWing (Complex Knowledge Base Tools for Natural Language Communication with the Semantic Web) project. The purpose of the project is to develop a complex base of tools which will remove some of the typical barriers present in communication between human user and computer within the scope of internet access. Our contribution includes methods, algorithms, tools and techniques improving the quality of user's experience during web usage, namely:

- voice communication including the making of inquiries in natural language,
- enrichment of query results based on recognition of user's area of interest,
- tools for automatic creation of semantic description of utterances in limited domains,
- tools for obtaining query results in different languages,
- means for web page and document filtration,
- methods for automatic annotation and summarization of large documents and document collections,
- acquisition of latent information and knowledge from web environment,
- tools for evaluation and search for domain experts, expert groups, authorities etc.

The system architecture and methods, algorithms, tools and techniques mentioned above are described in the following sections in detail.

2 System architecture

Figure 1 describes the architecture of the whole system for natural language communication with semantic web. The functioning of system components is described in the respective sections.

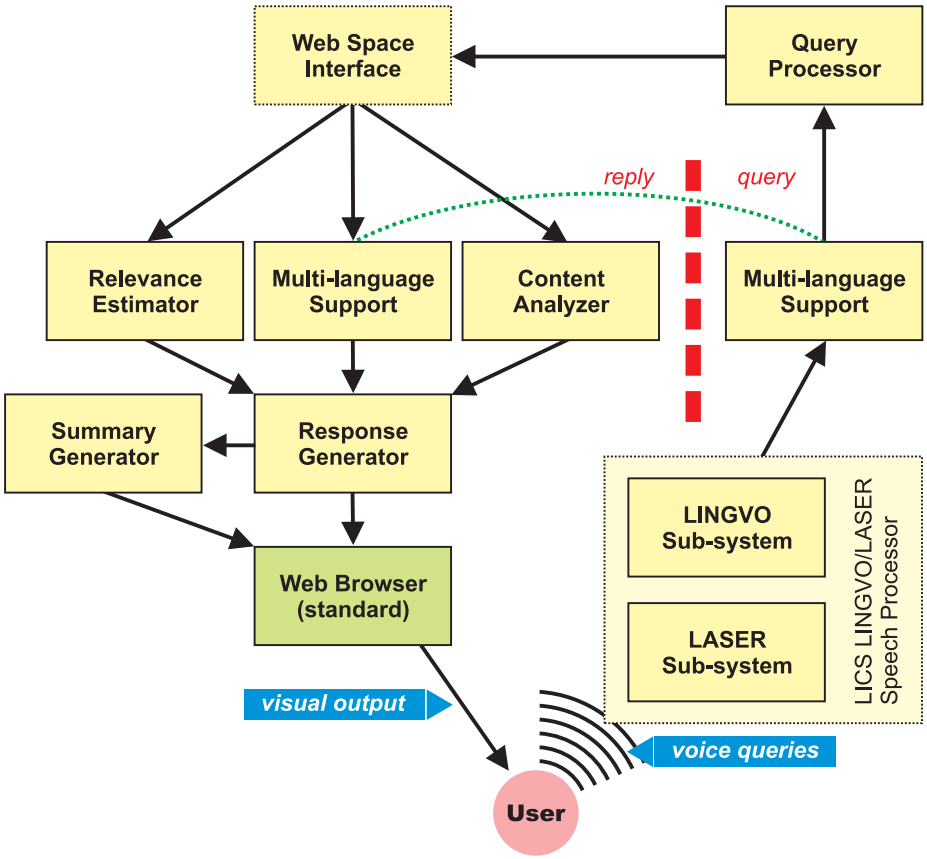

Fig. 1. System architecture diagram

3 Voice Communication Including the Making of Inquiries in Natural Language

The LASER (LICS Automatic Speech Extraction/Recognition) recognizer [EP04] is currently being developed by the Laboratory of Intelligent Information Systems (LICS), University of West Bohemia and will be used to transform the

queries and commands spoken in natural language. We are currently working on both traditional hidden Markov model based speech recognition as well as on a so called hybrid approach which tries to combine the advantages of hidden Markov models (HMMs) and artificial neural networks. Our first experiments (see e.g. [Pav03]) were carried out with HMMs describing all the utterances that the system recognizes. While this is suitable for simple tasks (i.e. small vocabulary and small number of distinct utternaces to recognize) with the growing complexity of the task the size of the HMM rises to a level where it places demands on computational power that are no longer feasible. In order to work with larger dictionaries and syntactically richer utterances it is necessary to:

- Find a way to generate the HMM "on the fly" during the recognition process.
- Discard (prune) those states that are unlikely to lead to the desired result.

These issues and their respective proposed solutions can be found in [Pav06].

Since it is not yet possible to recognize general speech (the best recognizers today can deal with dictionaries having about 200 thousand words which is not enough for inflectional languages like Czech) the biggest challenge is how to model the limited domain of internet queries.

To define the domain means to find (and model) a subset of all the possible sentences in the language that are sufficient to cover all commands and inquiries the user can make up. One possibility supported by our recognizer is the usage of context free grammars, but the disadvantage is that the grammars have to be made by hand. Another solution is to use stochastic language models which can learn the syntax of the language from a set of training examples. Where to find the training data is a question for further research.

4 Tools for Automatic Creation of Semantic Description of Utterances in Limited Domains

The first task within the building tools for automatic creation of semantic description in limited domains is determination of the goal domains and corpora preparation. We have decided to elaborate on domains, in which the answer to a simple question can be easily found in web pages. These domains include e.g. weather forecast, public transport, accommodation and food services, local authorities and government offices, on-line shopping, information about monuments and museums.

Building a corpus is a long-lasting and time-consuming process demanding a large number of participants. There is also a need to map the target group of possible users of the whole system and to find the way in which they would ask for information from the web. According to our short survey, only a very limited group of people using Internet search engines uses a sequence of keywords to find appropriate information as a standard way of thinking and expressing. Most people prefer to make inquiries in natural language. Hence, there is necessary to consider a making inquiry in natural language as a standard process of web information search.

We have collected about 27.000 questions in ten domains so far. The corpus has been built by approximately 450 people (half of them were students, half of them their relatives, friends etc.). All the participants created questions according to guidelines given them by our research group. Approximately 70% of collected questions can be used in the following process the semantic annotation.

The corpus of typical user utterances has to be annotated. The process of annotation is the process of assigning the meaning to each utterance. The meaning is assigned by a team of human annotators. The meaning of a sentence is represented in a suitable meaning representation. Hence, one step is to propose the suitable representation of the meaning. When the corpus is annotated (at least partially) then a computer algorithm has to be proposed to learn automatically the rules of semantic analysis. The result of this stage of the COT-SEWing project is an algorithm that is capable of doing automatic semantic analysis. Such an algorithm automatically assigns the meaning to an input utterance.

Current development of semantic analysis modules focus on stochastic learning. Our system is also based on a stochastic method [Kon06]. The bright side of stochastic systems is that they can easily deal with wide variety of nuances in input utterances. The problem of stochastic method are complicated phrases because they are too complicated to be learned automatically. And thus we use a preprocessing phase that identifies such complicated phrases and parses them before the main stochastic algorithm is executed.

5 Means for Web Page and Document Filtration

Internet is an immense resource of data. There are billions of documents in various formats text, image, audio, video, etc. Many of these documents represent useful knowledge that must be extracted out of them first. This extraction (mining) is the subject of a scientific field called Web mining. Recent advances in Web mining have concentrated on content, structure, and usage mining. Both content and structure mining techniques may help us distinguish between relevant and irrelevant Web documents. The former by categorizing into on-topic and off-topic documents, the latter by determining important (authoritative) documents via analysis of their relations to other documents. Of course, the heterogeneous and decentralized nature of the Web causes many useless, harmful or even criminal Web pages to appear. Web mining may also be applied to their detection and elimination.

When visiting a Web page on pornography, racism, genocide or criminality topics, it is often not very clear whether the page represents a violation of law or whether it is just a pure description of the domain without any hidden intentions. We often need a domain expert for an exact evaluation. Of course, there is a difficulty with accessibility of such pages as well. It is not only complicated to recognize them but also to find them. Some of the improper topics mentioned above are by far not as frequent as the others. It is a tedious task.

Therefore, at the first stage of our task we want to create manually a collection of specific textual documents definitely violating the law according to the

immoral offence article. This dataset will be then used for training a suitable text classifier which will help us to recognize other improper Web documents and collect them automatically. The goal is not only to create a more extensive dataset covering an improper topic, but also to extract other suitable information from such Web pages (e.g. frequently mentioned email addresses or names, servers containing many improper Web pages, geographic location of these servers, link interconnection between these pages etc.). This careful analysis can provide valuable data useable not only for other researchers and scientists, but also for institutions authorized to restrict the amount of problematic Web pages on the Internet.

To realize this goal we proposed the preliminary version of a Web crawling system which would be able to detect pages with a specific topic (or topics), analyze them and store various information about them. The structure of our system which has been named WebRunner is depicted in the Figure 2.

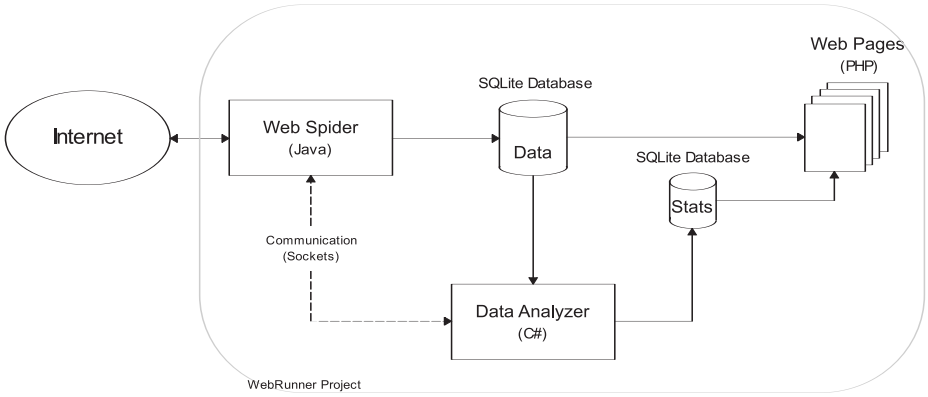

Fig. 2. Web Runner Project

We expect that the Web Spider module (see the picture above) will contain various sub-modules for real-time page analysis. Namely, it should be an HTML parser (able to extract links and email addresses), text parser, text classifier for topic detection, language recognizer, and a sub-module for geographic location of a given Web server. All the obtained data should be then stored in a database. We would like to utilize the SQLite database because of its simplicity and performance. The purpose of the Data Analyzer module will be the further analysis of stored data and various useful statistics generation (e.g. number of word occurrences in single pages, etc.). To provide public access to the data and statistics we plan to create an appropriate Web interface.

Thanks to the universality of this system not only improper, but also any other topics on the Internet can be analyzed and used to support other tasks of our project. For example, documents describing various natural disasters can

be recognized, stored, and used for further latent semantic analysis (see section 6.2).

Indeed, there are many hidden problems before us. To solve them it will be necessary to perform some experiments and to examine obtained results. Among the major questions are what text classifier should be used for Web page topic detection and what document model should be used to reach maximum accuracy of our text classifier. So far, we examined some possibilities how to extend the common widely used bag-of-words document model, which completely ignores conceptual patterns. Our preliminary results (see [TPSJ06]) indicate that word n-grams, beside others, could be suitable. Using only bag-of-words approach multi-word expressions with a special meaning like "United States" are chunked into pieces, sometimes with comletely different meaning like "united" and "states". And it can cause an unnecessary loss of accuracy. In our next step we would like to unite synonymous words (e.g. "sick-ill", "baby-infant"), which are generally treated like two different words and we would like to use word sense disambiguation as well to avoid problems caused by polysemous words (e.g. "train", "can"). This will be done using EWN (see section 7) and its semantic relations.

However, the generation of word n-grams can be sometimes difficult especially in the case of large datasets. Thus, on the basis of our observations in [TFRJ05], we also plan to propose a new algorithm for n-gram discovery from very large text datasets.

6 Methods for Automatic Annotation and Summarization of Large Documents and Document Collections

User's experience during web usage can be largely improved by text summarization. Sending a short summary instead of a list of documents as a result of a query would make the web really semantic. We plan to pursue both single- and multi-document summarization in multiple languages.

6.1 Languages, Corpora and Annotation

We would like to focus on various languages in summarization part of the project. We assume to start with English and Czech. As for testing corpora, we are going to use standard DUC corpora for English. And because there is no corpus annotated for summarization in Czech, we plan to create and annotate a new one. An XML corpus format will be developed and the corpora will be converted into it. Further, we will create an annotation tool that will enable to work with the XML format, to annotate the corresponding anaphoric expressions and sentence compression. As a result of the annotation we will get a corpus annotated for sentence extraction, anaphora resolution (cross-document coreference resolution in the case of multi-document summarization) and sentence compression.

6.2 Single-document Summarization

We developed a summarization approach that is based on latent semantic analysis (LSA) [SJ04]. LSA [LD97] is a technique for extracting the hidden dimensions of the semantic representation of terms, sentences, or documents, on the basis of their contextual use. The core of the analysis is the singular value decomposition (SVD) of the input documents terms-by-sentences matrix. The idea of our summarization approach is to identify the most important topics from the source text and then to choose the sentences with the greatest combined weights across the topics. Afterwards, we enriched the document representation by anaphoric relations [SKP05]. Anaphoric resolver GuiTAR [PK04] has been used for automatic annotation of anaphoric expressions. We are going to compare these annotations with those made by human by the annotation tool. It was found that the addition of anaphoric knowledge leads to improved performance of the summarizer. Later, we went beyond sentence extraction and proposed a simple sentence compression algorithm for our summarizer [SJ06]. We will try to improve the sentence compression and compared them with those made by human by the annotation tool. Summaries are used in our MUSE (Multilingual Searching and Extraction) system [TSJ06]. They enable better and faster user orientation in retrieved results.

6.3 Multi-document Summarization

We are currently working to apply the methods proposed here to multi-document summarization. The single-document LSA approach can be easily extended to process multiple documents by including all sentences in a cluster of documents in the SVD input matrix. The sentence selection approach can be used as well; however, care has to be taken to avoid including very similar sentences from different documents. Therefore, before including a sentence in the summary we have to check if there are any sentences whose similarity with the observed one is above a given threshold. (The easiest way of measuring the similarity between two sentences is to measure the cosine of the angle between them in the term space.) Cross-document coreference, on the other hand, is a fairly different task from within-document coreference, as even in the case of entities introduced using proper names one cannot always assume that the same object is intended, let alone in the case of entities introduced using definite descriptions. We are currently working on this problem. Sentence compression can be used in the same way as in the case of single-document approach.

6.4 Semantic Web Summarization System Vision

The cluster typically contains related documents (e.g., hurricane disaster). We plan to create an interface where a user could enter a query (e.g., casualty toll) and a query-focused summary extracted from documents in the cluster will be returned. In the end the cluster will be created from online documents returned by a web searching engine. This will make the system really semantic.

7 Tools for Obtaining Query Results in Different Languages

Multilingual aspects have been gaining more and more attention in recent years. This trend has been accentuated by the global integration and the vanishing cultural and social boundaries. The ever-increasing use of foreign languages is due to the information boom caused by the huge amount of web content. Multilingual text processing has become an important field bringing a lot of new and interesting problems. Particularly when a text corpus includes documents written in various languages, the way of their processing is not satisfactorily solved so far. We intend to focus our attention on them and their possible solutions are proposed in our project COT-SEWing. We suppose that a multilingual system will be useful in text-based tasks, as well as in the semantic web.

We intend to use the EuroWordNet thesaurus (EWN) as the core of our multilingual approach. Due to EWN it is possible to perform a language independent processing and transform the text into the language independent form. The internal format enables the creation of queries in various EWN languages. Within the scope of the work, we developed a searching system with a query expansion module that should lead to easier query creation. Thanks to the EWN it is possible to perform multilingual, monolingual as well as cross-language processing. We focus our attention mainly on Czech and English language.

EuroWordNet thesaurus can be applied in many Natural Language Processing (NLP) areas. It is a multilingual database of words and relations for most European languages. It contains sets of synonyms - synsets - and relations between them. It interconnects the languages through an inter-lingual-index in such a way that the same synset in one language has the same index in another one.

Thanks to the EWN-based approach, it is possible to perform additional techniques in the processing e.g. query expansion, cross-language information retrieval, and word sense disambiguation. We have created a multilingual searching and extraction system called MUSE [TSJ06]. It represents a prototype system to verify our EWN-based methods and it demonstrates possibilities, advantages, and disadvantages of our approach.

A query entered to the system (e.g. to the semantic web) can be expanded to obtain a broader set of results. EWN relations between synsets are used for the query expansion. Hypernym, holonym or other related synsets can enhance the query. According to our preliminary results a query expansion can significantly improve the system recall. It means that the system will retrieve more information, which is still relevant to the query. Moreover, thanks to EWN use we are able to perform cross-language information retrieval.

Word sense disambiguation is another NLP task leading to understanding the semantic meaning of the word in most of the natural language processing systems. It allows to distinguish the meaning of a text, a message, or a query. Polysemous words may occur in any language. Ambiguity causes many problems, which may result in the retrieval of irrelevant information. We propose to use the EWN thesaurus as a main solving tool. We intend to implement a disambiguation

method based on the Bayesian classifier. Each meaning of the word is represented by a class in the classification task. The total number of meanings for each ambiguous word is obtained from the EWN thesaurus.

According to preliminary tests, the multilingual semantic aware approach used in our prototype system gives promising results. We are able to use EWN thesaurus in many NLP tasks in such a way that the semantic meaning of the word is preserved and can be used in other processing tasks. Moreover, we could retrieve the results in different languages or simply perform multilingual and cross-lingual information retrieval and classification [TJ05].

8 Acquisition of Latent Information and Knowledge from Web Environment

Recently, it has become clear that studying the structure of the Web helps extract much useful information. The knowledge of structure is sometimes even more valuable than the knowledge of content. Since the Web is most often modeled as a graph, the importance of its structure (or topology) is indisputable. A common task, performed by Web search engines among others, is to determine importance of a Web site or Web page. This is often done by exploring its relations to other Web pages in terms of hyperlinks among pages in analogy to counting bibliographic citations in scientific literature. There are a couple of common ranking algorithms that assign quality rankings to Web pages, based on the structure of the Web graph [DHH$^+$02].

Our initial task, in scope of this project, is to determine authoritative institutions within a set of Czech academic computer science Web sites. We have chosen this area because we know it well and we expect that there will be enough data on the Web to analyze. However, the methodology we are currently developing is sufficiently general and it can be applied to a different domain as well. In a Web directory we plan to select a certain number of highly representative computer science departments. Unfortunately, what makes the selection a little bit complicated is the fact that not all Czech universities have the same structure and the same hierarchy of faculties and departments.

One condition that limits our selection of experimental Web sites is that each department should have its home page on a Web server of its own and not of its faculty or University. Separate servers can be more easily processed by automated Web agents because the robot immediately recognizes whether a link is internal (within the server) or external. Some well-known heuristics also say that longer URLs are less important documents than the shorter ones. So, from this point of view, we will firstly leave out less significant sites right from the start.

Moreover, there are some facts that might have a much larger impact. For instance, existence of server aliases is annoying. If there are two host names representing one machine with the same content, references to them should be counted together. There may be a large number of aliases and ignoring them may yield significantly distinct results. Another trouble is dynamically generated

Web pages. Two and more URLs (and thus two and more possible citations) may represent one document and citations by them should be counted only once then. This is painful, especially with regard to generally low inter-linkage among sites. There is also a difficulty with document formats. If the ignored documents are more frequent on one server than on the others, this host will be disfavored. Therefore, we must take into account all these possible effects before declaring the most authoritative sites.

9 Conclusion

In this article we have described system architecture proposal within the solution of COT-SEWing project. We have also introduced methods, algorithms, tools and techniques proposed or already used to improve natural language communication with semantic web. The full range of approaches described above is supposed to be developed further in our project.

Acknowledgements

This work was supported by grant no. 2C06009 Cot-Sewing.

References

[DHH+02] C. Ding, X. He, P. Husbands, H. Zha, and H. Simon. Pagerank, hits and a unified framework for link analysis. In *Proc. 25th ACMSIGIR Conference on Research and Development in Information Retrieval*, pages 353–354, 2002.

[EP04] K. Ekštein and T. Pavelka. Lingvo/laser: Prototyping concept of dialogue information system with spreading knowledge. In *NLUCS'04*, pages 159–168, Porto, Portugal, April 2004.

[Kon06] Miloslav Konopík. Stochastic semantic parsing. Technical Report DCSE/TR-2006-01, University of West Bohemia in Pilsen, 2006.

[LD97] T. K. Landauer and S. T. Dumais. A solution to plato's problem: The latent semantic analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2):211–240, 1997.

[Pav03] T. Pavelka. Hybrid speech recognizer implementation. Master's thesis, Universisty of West Bohemia, 2003.

[Pav06] T. Pavelka. Ldec: One pass time synchronous decoder. In *PhD Workshop 2006*, Hrub Skla, Czech Republic, October 2006.

[PK04] M. Poesio and M.A. Kabadjov. A general-purpose, off-the-shelf anaphora resolution module: implementation and preliminary evaluation. In *Proceedings of LREC*, 2004.

[SJ04] J. Steinberger and K. Jezek. Text summarization and singular value decomposition. In *Proceedings the 3rd International Conference on Advances in Information Systems, Lecture Notes in Computer Science 2457*, pages 245–254. Springer-Verlag, 2004.

[SJ06] J. Steinberger and K. Jezek. Sentence compression for the lsa-based summarizer. In *Proceedings of the 7th International Conference on Information Systems Implementation and Modelling*, 2006.

[SKP05] J. Steinberger, M. A. Kabadjov, and M. Poesio. Improving lsa-based summarization with anaphora resolution. In *Proceedings of Human Language Technology Conference/Conference on Empirical Methods in Natural Language Processing*, pages 1–8. The Association for Computational Linguistics, 2005.

[TFRJ05] R. Tesar, D. Fiala, F. Rousselot, and K. Jezek. A comparison of two algorithms for discovering repeated word sequences. In *The 6th International Conference on Data Mining, Text Mining and their Business Applications (Data Mining 2005)*, pages 121–131, 2005.

[TJ05] M. Toman and K. Jezek. Documents categorization in multilingual environment. In *Proceedings of the 9th International Conference on Electronic Publishing ELPUB2005*, 2005.

[TPSJ06] R. Tesar, M. Poesio, V. Strnad, and K. Jezek. Extending the single words-based document model: A comparison of bigrams and 2-itemsets. In *The 2006 ACM Symposium on Document Engineering (DocEng?06)*, pages 138–146. The Association for Computing Machinery, 2006.

[TSJ06] M. Toman, J. Steinberger, and K. Jezek. Searching and summarizing in multilingual environment. In *Proceedings of the 10th International Conference on Electronic Publishing*, 2006.